

A COMPUTATIONAL FRAMEWORK FOR THE VISUAL CORRESPONDENCE PROBLEM

Michael Kass

MIT Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139

Abstract

A computational framework for solving the visual correspondence problem is presented and evaluated by using a stochastic image model. The framework differs principally from previous work in that it emphasizes the combination of a large collection of independent measurements. Partial derivatives of images smoothed with a few different-sized Gaussian filters are suggested as suitable measurements. A specific computation is shown to reliably establish whether or not two points correspond, provided that the signal to correspondence noise ratio in the images to be matched exceeds two. The computation has been applied to artificial and natural images with encouraging results.

The problem of matching up two similar views of the same scene is one of the critical problems which any powerful vision system must solve. Known as the correspondence problem, it occurs most notably in stereopsis where the two views come from separate vantage points, and in motion analysis where the two views come from the same vantage point but are separated in time. In stereopsis, a solution to the correspondence problem yields relative depth information fairly directly. In motion analysis the correspondence is more difficult to interpret, but it includes information about the velocities and 3-dimensional structure of the objects being viewed. The human visual system is known to solve both correspondence problems with impressive range, resolution, and noise immunity though the manner in which it does so is ill understood. Computer solutions to the correspondence problem have fallen far short of similar performance, particularly on natural images. A new approach to the problem will be presented here which, it is hoped, will provide insight into the structure of the correspondence problem, as well as a robust technique for solving it.

The correspondence problem can be stated quite simply as follows. Given two similar images of the same scene, a point in the first image is said to correspond with a point in the second image if both are projections along lines of sight of the same physical point. The correspondence problem consists of trying to match up as many pairs of corresponding points as possible given the intensity profiles of the two similar images.

All algorithmic solutions to the problem are based on the idea that the light intensity profiles surrounding corresponding points are quite similar. For each point in the first image, only points in the second image with quite similar local intensity profiles need be considered as potential matches. If the similarity measure is chosen appropriately, then a large fraction of the points in the first image will have only one potential match in the second. If it can be confidently determined that the similarity between these points and their potential matches is not due to chance, then the unique potential matches can be trusted as correct matches. Global consistency constraints can be used in some cases to choose among several potential matches,

but this may not be necessary if the local information is extracted properly.

Previous work on the correspondence problem has centered on two types of similarity measures. The most straightforward algorithms have used similarity measures such as correlation or mean square intensity differences [3] [10] while more recent algorithms have relied on measures which declare two points to be similar if they fall on edges of approximately the same orientation [1] [2] [4] [9]. Neither approach has yet achieved the kind of noise immunity needed for a truly practical solution to the correspondence problem.

This paper will explore the possibility of increasing noise immunity by combining independent measures of similarity. No single similarity measure will be trusted much in isolation, but unanimity among the measures will be taken as a strong indication of correspondence. Using this approach, algorithms can be constructed which will tolerate signal-to-noise ratios as low as two to one, while localizing matches to within a pixel.

Let $I_1(x, y)$ and $I_2(x, y)$ be the light intensity functions for two images whose correspondence is to be computed and let $D(x, y)$ be the true offset or disparity between the images measured relative to the coordinate system of $I_1(x, y)$ and defined on some set of points $\mathcal{D} \subseteq \mathbb{R}^2$ for which corresponding physical points are visible in both images. Then for all $p \in \mathcal{D}$, $I_1(p)$ and $I_2(p + D(p))$ are projections of the same physical point. The problem is to recover $\mathcal{D}$ from I_1 and I_2 .

Consider a collection $f_i(p, I)$, $1 < i < n$ of different image functionals (filters). Each functional implicitly defines a similarity measure for correspondence since we expect that $f_i(p, I_1) \approx f_i(p + D(p), I_2)$ provided the f_i are chosen carefully. If we combine the functionals into a vector at each point: $F(p, I) = (f_1(p, I), f_2(p, I), \dots, f_n(p, I))$ then we can expect the vector $F(p_1, I_1) - F(p_2, I_2)$ to be very small in each component if p_1 and p_2 correspond. On the other hand, if p_1 and p_2 do not correspond, it is likely that $F(p_1, I_1) - F(p_2, I_2)$ has at least one large component.

The above intuition can be translated into an algorithm as follows. Define $matchp_i(p_1, p_2)$ be a predicate which is true if and only if

$$|f_i(p_1, I_1) - f_i(p_2, I_2)| < k_i \sigma(f_i(p, I_1))$$

where $\sigma(x)$ denotes the square root of the expected value of x^2 and let $matchp(p_1, p_2)$ be a predicate which is true if and only if for all $i \in \{1, 2, \dots, n\}$, $matchp_i(p_1, p_2)$. Then $matchp$ is true of a pair of points p_1 and p_2 if each component of $F(p_1, I_1) - F(p_2, I_2)$ is smaller than its globally determined threshold. It will be argued that $matchp$ does a good job of solving the correspondence problem if the f_i and the k_i are chosen appropriately—it is almost always true of corresponding points and almost never true of non-corresponding points.

In order to evaluate $matchp$, suppose the f_i are orthogonal linear shift invariant functionals and consider the following stationary image model. Let I_1 be stationary Gaussian white noise and let I_2 be derived from I_1 by shifting it according to $D(p)$ and adding Gaussian white correspondence noise, $N(p)$. The efficacy with which $D(x, y)$

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's Artificial Intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643, the Office of Naval Research under contract number N0014-80-C-0505, and the System Development Foundation.

can be determined from the f_i under these conditions depends upon how well the f_i are preserved between views. Let

$$SNR_i = \sigma(f_i(p, I_1)) / \sigma(f_i(p, N))$$

be the signal-to-correspondence-noise ratio of the i th functional. If SNR_i is greater than two for a dozen functionals, then *matchp* will very reliably determine whether or not two points correspond.

Three performance criteria will be considered for *matchp*. The first is the rate of false negatives—the probability that *matchp* will be false of two corresponding points. The second criterion is the rate of false positives—the probability that *matchp* will be true of two non-corresponding points. The third criterion is one of resolution and concerns the extent to which corresponding points can be spatially localized.

The false negative rate is [6] [7]

$$Pr[\sim \text{matchp}(p_1, C(p_1))] = 1 - \prod_{i=1}^n \operatorname{erf}\left(\frac{\sqrt{2}}{2} k_i SNR_i\right).$$

The false positive rate is [6] [7]

$$Pr[\text{matchp}(p_1, p_2)] \approx \prod_{i=1}^n \operatorname{erf}\left(\frac{k_i}{2}\right)$$

Suppose $SNR_i = 2$ for all i and $n = 12$. Then the choice of k_i represents a tradeoff between a very low false positive rate and a very low false negative rate. False positives often result in the generation of wrong disparity values so they tend to be quite serious. False negatives, on the other hand, usually result simply in not being able to determine the disparity at a particular point. Thus a reasonable choice of k_i is one which produces a negligible false positive rate while still keeping the false negative rate to a low level. One such choice is $k_i = 1.2$. The resulting false positive rate is .2 per cent and the resulting false negative rate is 18 per cent. Both rates are a good deal lower than what is needed to reliably determine image correspondence. If the signal-to-correspondence-noise ratio is improved to three, the false positive rate can be improved an order of magnitude without worsening the false negative rate.

The third criterion of performance for *matchp* is that of resolution. If *matchp*(p_1, p_2) is true, then *matchp*($p_1, p_2 + r$) is quite likely to be true if r is small enough. The value r_s of r at which the probability of *matchp*($p_1, p_2 + r$) drops to fifty percent is determined primarily by the functional with the most sharply peaked autocorrelation function. If A_i is the autocorrelation (normalized to be one at zero) of the i th functional, the resolution can be approximated by [6] [7]:

$$r_s \approx \inf_i \frac{3}{2} k_i \left(\frac{\partial^2 A_i}{\partial r^2} \right)^{-1/2} \Big|_0$$

In deriving the properties of *matchp* which allow it to be used to solve the correspondence problem, the existence of a set of independent, linear shift invariant functionals whose values are loosely preserved between views was assumed. One such set will be presented here. If p_1 is a point in $\mathcal{D}$ and p_2 is its corresponding point then the functions $I_1(p_1 + r)$ and $I_2(p_1 + D(p_1) + r)$ can be expected to be quite similar for small values of r . One complete characterization of the local behavior of a function of two dimensions is its two dimensional Taylor series, so it is natural to examine derivatives of $I_1(p_1 + r)$ and $I_2(C(p_1) + r)$. As one might expect, first and second partial derivatives appear empirically to be fairly well preserved between views. Differentiation tends to accentuate noise, so it is usually a good idea to do some low pass filtering before taking any sort of derivative. Marr and Hildreth [5] [8] argue that the best low pass

filter to use for applications such as this is a filter with a Gaussian point-spread function because it minimizes the product of localization in space and frequency. Hence a reasonable set of functionals to look at is the set of derivatives of Gaussian smoothed images.

Not all derivatives of Gaussian smoothed images are independent. In fact, the n th and $n+2$ nd derivatives of Gaussian smoothed white noise are very strongly correlated. These correlations can be calculated fairly directly. For notational convenience, let

$$f_{n,m,c}(x,y) = \frac{\partial^n}{\partial x^n} \frac{\partial^m}{\partial y^m} \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

Then the correlation between f_{n_1,m_1,σ_1} and f_{n_2,m_2,σ_2} is [6] [7]

$$-1^{n_1+m_1} \left(\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2} \right)^{m+n+2} \frac{n!m!}{(n/2)!(m/2)!} \sqrt{\frac{n_1!m_1!n_2!m_2!}{(2n_1)!(2m_1)!(2n_2)!(2m_2)!}}$$

For any particular amount of Gaussian smoothing, the largest sets of independent derivatives are of cardinality four. One such is the following:

$$\mathcal{F}_\sigma = \left\{ \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2} \right\}$$

If the ratio between the standard deviations of the two Gaussians is $s = \sigma_1/\sigma_2$ then the effect of the size of the Gaussians on their correlation can be expressed by the relation:

$$\operatorname{Corr}(f_{n_1,m_1,\sigma_1}, f_{n_2,m_2,\sigma_2}) = \left(\frac{2s}{s^2 + 1} \right)^{m+n+2} \operatorname{Corr}(f_{n_1,m_1}, f_{n_2,m_2})$$

Let $\mathcal{F}^* = \mathcal{F}_\sigma \cup \mathcal{F}_{\sigma s} \cup \mathcal{F}_{\sigma/s}$. The maximum correlation between two functionals in $\mathcal{F}^*$ is $(2s/(s^2 + 1))^4$ which occurs between first order terms. If $s = 2$, the correlation is .41 but if $s = 2.5$, it drops to .23 and if $s = 3$, it falls to .13. The impact of these small, non-zero correlations on the performance of *matchp* can safely be ignored.

As an initial test, *matchp* was applied to a pair of gray level images generated by computer with all the important characteristics of the image model used here. $\mathcal{F}^*$ was used as the set of functionals. The image pair has a signal-to-correspondence-noise ratio of two and a disparity field which is zero everywhere except in a central square covering one ninth of the image area where it is (6,0) in pixels. For each point (x,y) in the left image, *matchp*($x+n,y$), $n \in \{-8,-7,-6,\dots,6,7,8\}$ was calculated. If there was only one n such that *matchp*($x+n,y$) was true, n was recorded as the disparity value. If there was more than one n such that *matchp*($x+n,y$) was true, the disparity was recorded as ambiguous. If there was no n such that *matchp*($x+n,y$) was true, the disparity was recorded as unknown.

Figure one shows the results of applying *matchp* to the above pair of images. The dark points indicate areas where the algorithm was unable to find matches. Slightly over 93 percent of the pixels in each image were uniquely matched. The mean square error in the disparity values generated was a small fraction of a pixel despite the large amount of correspondence noise.

It is worth noting that *matchp* failed to match most of the points along the border of the shifted central square. Some of the points were occluded in the second image, so it was correct not to match them, but most of the points went unmatched because the steep disparity gradient on the border substantially decreased the signal-to-noise ratio. Correlation and edge-based algorithms tend to generate significant numbers of incorrect disparity values at occluded regions

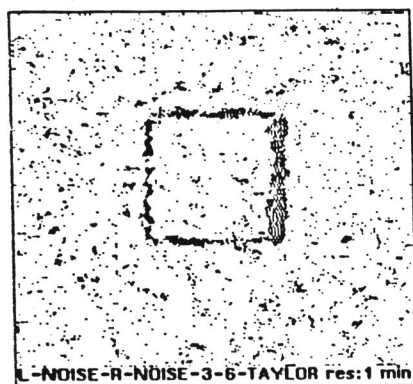


Figure 1. *Matchp* applied to a Julesz random dot stereogram. Dark points were unmatched.

and at places where the disparity gradient is large, but the algorithm based on *matchp* avoids doing so because of *matchp*'s unusually low false positive rate.

Matchp has been applied to a small number of natural images as well. Figure 3 shows the results of interpolating a surface through disparity values generated by *matchp* for the stereo pair in figure 2. Intensity is proportional to depth. The photos are of the campus of the University of British Columbia and were obtained from the B.C. Ministry of Forests.

The combination of independent results has long been a favorite method of statisticians. *Matchp* represents an attempt to bring the power of this method to bear on the visual correspondence problem. Despite using the simplest method of combination imaginable, *matchp* attains a rather high level of performance and so argues strongly for the applicability of this statistical tool to the correspondence problem.

Acknowledgements

Tomaso Poggio and Keith Nishihara provided important guidance. R.J. Woodham and Eric Grimson made the stereo photographs available.

References

- [1] Baker, H.H. and Binford, T.O. *Depth from edge and intensity based stereo* Proceedings of the Seventh International Joint Conference on Artificial Intelligence, 1981, 631-636.
- [2] Baker, H.H. *Depth from edge and intensity based stereo* Stanford Artificial Intelligence Laboratory Memo AIM 347, 1982.
- [3] Gennery, D.B. *A stereo vision system for an autonomous vehicle* Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 1977, 576-582.
- [4] Grimson, W.E.L. *From images to surfaces* M.I.T. Press, Cambridge, Mass., 1981.
- [5] Hildreth, E. C. *Implementation of a theory of edge detection* S.M. Thesis, Department of Computer Science and Electrical Engineering, M.I.T., 1980.
- [6] Kass, M. *An intensity based approach to the visual correspondence problem* B.A. Thesis. Department of Independent Concentration, Princeton, 1982.
- [7] Kass, M. *Computing Visual Correspondence* Proceedings: Image Understanding Workshop, June 1983.
- [8] Marr, D. and Hildreth, E.C. *Theory of edge detection* Proceedings of the Royal Society of London B 207, 1980, 187-217.
- [9] Marr, D. and Poggio, T. *A theory of human stereo vision*, Proceedings of the Royal Society of London B 204, 1979, 301-328.
- [10] Moravec, H.P. *Towards automatic visual obstacle avoidance*, Proceedings of the Fifth International Joint Conference on Artificial Intelligence

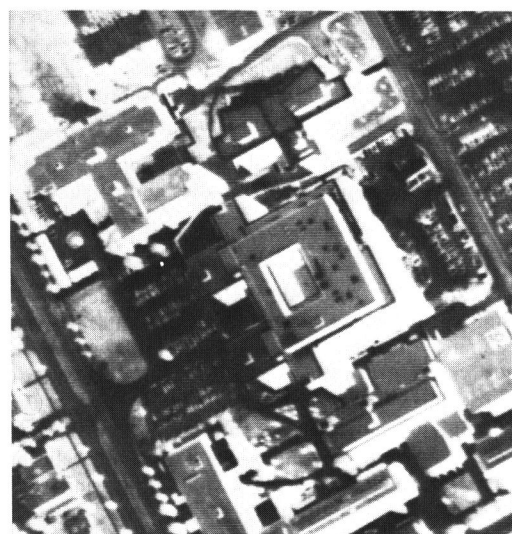


Figure 2. University of British Columbia from the air

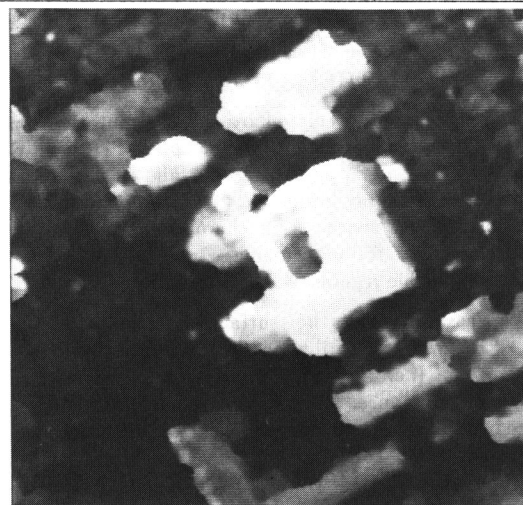


Figure 3. *Matchp* output. Intensity is proportional to depth.